ORIGINAL ARTICLES

Imputing missing standard deviations in meta-analyses can provide accurate results

Toshi A Furukawa*, Corrado Barbui, Andrea Cipriani, Paolo Brambilla, Norio Watanabe

*Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-cho, Mizuho-ku, Nagoya 467-8601 Japan
bDepartment of Medicine and Public Health, Section of Psychiatry, University of Verona, Ospedale Policlinico, 37134 Verona, Italy
cDepartment of Pathology and Experimental and Clinical Medicine, Section of Psychiatry, University of Udine, Policlinico Universitario via Colugna 50, 33100 Udine, Udine, Italy
dHealth Services Research Department, Institute of Psychiatry, King’s College London, University of London, De Crespigny Park, London SE5 8AF, UK

Accepted 16 June 2005

Abstract

Background and Objectives: Many reports of randomized controlled trials (RCTs) fail to provide standard deviations (SDs) of their continuous outcome measures. Some meta-analysts substitute them by those reported in other studies, either from another meta-analysis or from other studies in the same meta-analysis. But the validity of such practices has never been empirically examined.

Methods: We compared the actual standardized mean difference (SMD) of individual RCTs and the meta-analytically pooled SMD of all RCTs against those based on the above-mentioned two imputation methods in two meta-analyses of antidepressants.

Results: Two meta-analyses included 39 RCTs of fluoxetine (n = 3,681) and 25 RCTs of amitriptyline (n = 1,832), which had actually reported means and SDs of the Hamilton Rating Scale for Depression. According to either of the two proposed imputation methods, the agreement between actual SMDs and imputed SMDs for individual RCTs was very good with ANOVA intraclass correlation coefficients between 0.61 and 0.97. The agreement between the actual pooled SMD and the imputed one was even better, with minimal differences in both their point estimates and 95% confidence intervals.

Conclusion: For a systematic review where some of the identified trials do not report SDs, it appears safe to borrow SDs from other studies.

Keywords: Meta-analysis; Standard deviation; Missing data; Imputation; Depressive disorder

1. Introduction

Conduct of a systematic review or a meta-analysis involves comprehensive search of relevant randomized controlled trials (RCTs) and their quantitative or qualitative synthesis. To pool results on a continuous outcome measure of the identified RCTs quantitatively, one needs both means and standard deviations (SDs) on that outcome measure for each RCT.

Many reports of RCTs, however, fail to provide SDs for their continuous outcomes. It is sometimes possible to use P or t or F values, reported in the original RCTs, to calculate exact SDs [1]. When none of these is available, it is recommended that one should contact primary authors [2]. However, the yield is very often very low; some are incontactable, some never respond, and others report that the data are discarded, lost or irretrievable because there are no longer any computers to read the tapes.

Some meta-analysts then resort to substitution of SDs of known outcome measures by those reported in other studies, either from another meta-analysis or from other studies in the same meta-analysis. But the validity of such practices has never been empirically examined.

The present article therefore aims to examine empirically the validity of borrowing SDs from other studies when individual RCTs fail to report SDs in a meta-analysis, by simulating the above-mentioned two imputation methods for SDs in two meta-analyses on antidepressants that we have conducted [3,4]. Systematic reviews for depression are particularly suitable for this purpose, because Hamilton Rating Scale for Depression [5] (HRSD) is the de facto standard in symptom assessment and is used in many depression trials identified for overviews.
2. Methods

2.1. Imputation from a previous meta-analysis

We used the pooled SD for the 17-item HRSD and 21-item HRSD from a comprehensive meta-analysis of all three-armed trials for depression involving an investigational drug, an active comparison drug, and placebo [6] for two reasons. We are unaware of any other meta-analysis that dealt with the whole range of antidepressants along with placebo for depression. We argue that such a meta-analysis would provide a more impartial estimate of SDs for ratings scales than, for example, a meta-analysis dealing with a particular agent. We also had access to the raw data of this meta-analysis, so that we could easily calculate the pooled SDs for each version of the HRSD separately, using the formula below:

$$SD_{\text{pooled}} = \sqrt{\frac{\sum (n_i-1)SD_i^2}{\sum (n_i-1)}}$$

The pooled SD for the 17-item HRSD was 7.6, based on six studies involving 1,732 patients. That for the 21-item HRSD was 8.2, based on eight studies involving 2,129 subjects.

2.2. Imputation from the same meta-analysis

We simulated the situation where meta-analysts would impute the missing SDs for a small number of trials from all the other studies in the same meta-analysis by way of “leaving-one-out” method [7]. Namely, we imputed SD for each trial by calculating the pooled SD from all the other studies in the same meta-analysis that used the same version of HRSD (i.e., either 17-item or 21-item versions), and then meta-analyzed these imputed effect sizes.

2.3. Procedures

We applied these two imputation methods to a comprehensive meta-analysis of fluoxetine, the most widely prescribed selective serotonin-reuptake inhibitor antidepressant [4], and another focusing on amitriptyline, a classic tricyclic antidepressant [3], and made the following comparisons between the actual values and the borrowed/imputed values.

1. Standardized mean difference (SMD) of individual RCTs, where SMD for the $j$th study is calculated according to the following formula:

$$\text{SMD}_j = \frac{\text{Mean}_{a,j} - \text{Mean}_{c,j}}{\text{SD}_j}$$

Mean$_a$ is the end-point mean value of the HRSD for the control group, and Mean$_c$ is that for the active group.

2. Pooled SMD according to the random effects model [8,9].

The concordance between individual SMDs was examined by use of ANOVA intraclass correlation coefficient (ICC) [10,11]. The random effects model SMD was calculated by The Cochrane Collaboration’s meta-analysis software, Review Manager 4.2 [12]. We also pooled imputed SMDs according to the fixed effect model but the results were almost identical and we report herein the results according to the random effects model only.

3. Results

3.1. Characteristics of individual RCTs in the two meta-analyses

Of 133 RCTs pooled to examine the overall efficacy for fluoxetine vs. other antidepressants, 108 used HRSD, 8 used Montgomery-Asberg Depression Rating Scale (MADRS) [13], and 14 used other scales. Of the first group, 36 used 17-item HRSD, 39 used 21-item HRSD, and the remaining 33 used other or unknown versions of HRSD. Moreover, of the 36 studies that used HRSD-17, only 24 reported SDs; of the 39 which used HRSD-21, only 15 reported SDs. We therefore compared the meta-analytic results for 39 RCTs ($n = 3,681$), which used either 17-item or 21-item HRSD and reported SDs. The reported SDs for HRSD-17 ranged between 4.10 and 10.00 (mean = 6.36), and those for HRSD-21 between 2.75 and 10.98 (mean = 8.15).

Of 52 RCTs that had been pooled to examine the overall efficacy of amitriptyline vs. other antidepressants and that had reported SDs for their outcome measures, 47 used HRSD, 3 used MADRS, and 2 used ad hoc scales. Of the first group, 16 used the 17-item HRSD, 9 used the 21-item HRSD, and the remaining 22 used unspecified or other versions of HRSD. We therefore compared the meta-analytic results for 25 RCTs ($n = 1,832$), which used either 17-item or 21-item HRSD to examine the effectiveness of amitriptyline. The reported SDs for HRSD-17 ranged between 5.85 and 14.76 (mean = 8.58), and those for HRSD-21 between 4.96 and 9.84 (mean = 7.40).

3.2. Results for the fluoxetine review

When we used the pooled SD from a previous meta-analysis [6], the concordance between actual SMDs and imputed SMDs was an ANOVA ICC of 0.96 (95% confidence interval: 0.93 to 0.98) (Fig. 1a). The pooled SMD based on actual values of the 39 RCTs was 0.10 (95% confidence interval: 0.03 to 0.16), whereas that from imputed values was 0.10 (0.04 to 0.17).

When we substituted the SD for each trial by that pooled from all the other trials in the same meta-analysis, ANOVA ICC between actual SMDs and imputed SMDs was 0.97 (0.94 to 0.98) (Fig. 1b). By pooling all these imputed SMDs, we obtained the imputed pooled SMD of 0.10 (0.04 to 0.17) for fluoxetine.
3.3. Results for the amitriptyline review

By using the pooled SD from a previous meta-analysis [6], the ANOVA ICC between actual individual SMDs and imputed individual SMDs was 0.94 (0.86 to 0.97) (Fig. 2a). The actual pooled SMD from the 25 RCTs was 0.18 (0.08 to 0.27), whereas the imputed pooled SMD was 0.18 (0.09 to 0.27).

By replacing the SD for each trial by that pooled from all the other trials in the same meta-analysis, the agreement between actual individual SMDs and imputed individual SMDs was an ANOVA ICC of 0.61 (0.29 to 0.80) (Fig. 2b). Nonetheless, the concordance between the meta-analyzed SMD was very good again, because the actual SMD was 0.18 (0.08 to 0.27) and the imputed SMD was 0.18 (0.09 to 0.27).

4. Discussion

Few studies to date have dealt with the problem of missing variance estimates for continuous variables in meta-analyses, although apparently the problem is annoyingly common. For example, in a systematic review of sodium reduction on blood pressure, fewer than half of the identified trials (10 out of 26 identified trials) published a variance estimate or information to allow one to derive it [14]. In the comprehensive meta-analysis of three-armed studies involving antidepressants [6], the situation was analogous, as only 14 out of 37 trials employing HRSD had reported SDs. In our fluoxetine review [4], barely half of the identified trials (39 out of 75 trials employing HRSD) had reported SDs or other information to allow exact calculation. The recent concerted efforts by journal editors and clinical epidemiologists to call for better reporting of trial findings represent a step forward to remedy this situation [15].
For today’s systematic reviews, however, ignoring studies without variance estimates appears to be common [16]. Unfortunately, not only does such a practice defeat the primary purpose of a systematic review to amalgamate all the available evidence to derive the most narrow, precise estimate possible of a treatment effect [17], but also may lead to a biased point estimate of a treatment effect as studies with missing variance information may well not be a random subset of all the available studies.

The usually advised solution to this problem is to ask primary authors for information but the effort is often fruitless [18].

An alternative solution then is to borrow SDs from other studies, either from a previous meta-analysis or from other trials in the same meta-analysis. We simulated these two methods in two large systematic reviews of antidepressants for depression by hypothesizing that the SDs retrieved from original studies had not been reported. The degree of concordance of the actual effect sizes and the imputed effect sizes was gratifying both on individual trial basis and on aggregate basis.

Strictly speaking, it is not straightforward to generalize the current findings beyond pharmacologic trials for depression with regard to the Hamilton Rating Scale for Depression.

However, the good to excellent correspondence between the actual SMDs and imputed SMDs of individual RCTs, and the virtual agreement between the actual meta-analyzed SMDs and the imputed meta-analyzed SMDs strongly argue for the appropriateness of both imputation methods. One must also remember that the present simulation study borrowing SDs from a previous meta-analysis represents the worst-case scenario, where none of the included trials had reported SDs, and therefore, the observed discrepancy, if any, would correspond with the biggest difference possible. In actuality, at least some of the identified trials do report SDs, and the resultant pooled estimates of the SMD would be less subject to the imputation assumption. Leaving out, for example, five of the included trials would be closer to borrowing from a different meta-analysis than the leaving-one-out method, which we employed in this article, but we felt that we did not need to simulate the former, as we had already examined the “worst case.”

At the moment we do not have much ground to choose between the two imputation methods. We would, therefore, like to recommend, in the case of systematic reviews where some of the identified trials do not report SDs:

1. When the number of RCTs with missing SDs is small and when the total number of RCTs is large, to use the pooled SDs from all the other available RCTs in the same meta-analysis. It is possible and recommended in this case to examine the appropriateness of the imputation by comparing the SMDs of those trials that had reported SDs against the hypothetical SMDs of the same trials based on the imputed SDs. If they converge, we can be more confident in the meta-analytic results.

2. When the number of RCTs with missing SDs is large or when the total number of RCTs is small, to borrow SDs from a previous systematic review, because the small sample size may allow unexpected deviation due to chance. One must remember, however, that the credibility of the meta-analytic findings will be less secure in this instance.

References